ESD-TR-69-63

Technical Note

1969-19

A. V. Oppenheim

Block-Floating-Point Realization of Digital Filters

20 March 1969

Prepared under Electronic Systems Division Contract AF 19(628)-5167 by

Lincoln Laboratory
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Lexington, Massachusetts

The work reported in this document was performed at Lincoln Laboratory, a center for research operated by Massachusetts Institute of Technology, with the support of the U.S. Air Force under Contract AF 19(628)-5167.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

LINCOLN LABORATORY

BLOCK-FLOATING-POINT REALIZATION OF DIGITAL FILTERS

A. V. OPPENHEIM

Group 62

TECHNICAL NOTE 1969-19

20 MARCH 1969

LEXINGTON MASSACHUSETTS

ABSTRACT

A realization for digital filters using block-floating-point arithmetic is proposed. A statistical model for roundoff noise is presented and used to compare block-floating-point with fixed-point and floating-point realizations.

Accepted for the Air Force
Franklin C. Hudson
Chief, Lincoln Laboratory Office

BLOCK-FLOATING-POINT REALIZATION OF DIGITAL FILTERS

Recently, the realization of digital filters by means of fixed-point and floating-point arithmetic have been compared on the basis of roundoff noise.[1] In this note, an alternative realization called block-floating-point is proposed. In block-floating-point arithmetic the input and filter states (i.e. the inputs to the delay registers) are jointly normalized before the multiplications and adds are performed with fixed-point arithmetic. The scale factor obtained during the normalization is then applied to the final output to produce a fixed-point output. To illustrate, consider a first-order filter described by the difference equation

$$y_n = x_n + a_1 y_{n-1} \tag{1}$$

To perform the computation in a block-floating-point manner, we define

$$A_n = \frac{1}{\mathcal{LP}\lceil \max\{|x_n|, |y_{n-1}|\}\rceil} \tag{2}$$

where $\mathcal{LP}\lceil M\rceil$ is used to denote the largest integer power of 2 which is less than or equal to M. Thus, A_n represents the power-of-two scaling which will jointly normalize x_n and y_{n-1}. We may then write y_n as

$$y_n = \frac{1}{A_n}\left[A_n x_n + a_1 A_n y_{n-1}\right] \tag{3}$$

or alternatively as either

$$y_n = \frac{1}{A_n}\left[A_n x_n + a\left(\frac{A_n}{A_{n-1}}\right) A_{n-1} y_{n-1}\right] \tag{4}$$

or

$$y_n = \frac{1}{A_n}\left[\left(\frac{A_n}{A_{n-1}}\right) A_{n-1}\, x_n + a\left(\frac{A_n}{A_{n-1}}\right) A_{n-1}\, y_{n-1}\right] \tag{5}$$

The representation of (4) is preferable to (3) since (3) implies that y_{n-1} is stored in the delay while (4) implies that $A_{n-1}\, y_{n-1}$ is stored in the delay. Since A_n is always greater than or equal to unity, $A_{n-1}\, y_{n-1}$ is represented more accurately than y_{n-1}. A disadvantage with the representation of (4) is that y_{n-1} must first be obtained to compute A_n, and $\left(A_n/A_{n-1}\right)$ must then be obtained. Equation (5) represents an alternative. Specifically, we note that

$$\frac{A_n}{A_{n-1}} = \frac{1}{\mathcal{QP}\left[\max\left\{\left|A_{n-1}\, x_n\right|, \left|A_{n-1}\, y_{n-1}\right|\right\}\right]} \tag{6}$$

Consequently, if we first scale x_n by A_{n-1} then the incremental scaling can be determined as specified by (6). If we consider the general case of an N^{th} order filter of the form

$$y_n = x_n + a_1\, y_{n-1} + a_2\, y_{n-2} + \cdots + a_N\, y_{n-N} \tag{7}$$

then the block-floating-point realization corresponding to (5) and represented in the direct form is depicted in Fig. 1. For the general case

$$\Delta_n = \frac{1}{\mathcal{QP}\left[\max\left\{\left|x_n\right|, \left|w_{1n}\right|, \left|w_{2n}\right|, \ldots, \left|w_{Nn}\right|\right\}\right]} \tag{8}$$

and

$$A_n = \frac{1}{\mathcal{QP}\left[\max\left\{\left|x_n\right|, \left|y_{n-1}\right|, \left|y_{n-2}\right|, \ldots, \left|y_{n-N}\right|\right\}\right]} = A_{n-1}\, \Delta_n \tag{9}$$

In evaluating the performance of the block-floating-point realization in the presence of roundoff noise we will restrict attention to first and second order filters. Furthermore we assume in the analysis that A_n is not constrained to be a scaling by a power of two. Finally, we assume that for the first and second order case one bit will be provided in the output register of the adder for overflow. This will always be sufficient for the first order filter, and is taken to be sufficient in a practical sense for the second order filter. Therefore, for the purpose of analysis we replace (8) and (9) by

$$\Delta_n = \frac{1}{2 \max \{ |x_n|, |w_{1n}| w_{2n}|, \ldots, |w_{Nn}| \}} \tag{10}$$

and

$$A_n = \frac{1}{2 \max \{ |x_n|, |y_{n-1}|, |y_{n-2}|, \ldots, |y_{n-N}| \}} \tag{11}$$

In the case of a first order filter a roundoff noise source is introduced in the multiplication by Δ_n, the multiplication by a_1, and the multiplication by $\frac{1}{A_n}$. Denoting these noise sources by ϵ_{1n}, ϵ_{2n} and ϵ_{3n} respectively, the resulting output noise η_n is, from (5),

$$\eta_n = \frac{1}{A_n} (\epsilon_{1n} + \epsilon_{2n}) + \epsilon_{3n} + a_1 \eta_{n-1}$$

Assuming that ϵ_{1n}, ϵ_{2n} and ϵ_{3n} are independent from sample to sample, and are independent of each other and $\frac{1}{A_n}$, then $\overline{\eta} = 0$ and

$$\overline{\eta^2} = \left(\overline{\epsilon_1^2} + \overline{\epsilon_2^2} \right) \frac{k_1}{1-a_1^2} + \overline{\epsilon_3^2} \tag{12}$$

where k_1 is the expected value of $\left(\frac{1}{A_n}\right)^2$ as specified by (11). In a similar manner, for the second order case, there are five noise generators as depicted in Fig. 2. Assuming that the noise generators are white, and independent of each other and A_n, and that all the noise generators have variance σ_ϵ^2,

$$\overline{\eta^2} = \sigma_\epsilon^2 + \sigma_\epsilon^2 (2+4r^2\cos^2\theta+2r^4)\, k_2 G \tag{13a}$$

where k_2 is the expected value of $\left(\frac{1}{A_n}\right)^2$ and G is given by

$$G = \left(\frac{1+r^2}{1-r^2}\right)\frac{1}{1+r^4-4r^2\cos^2\theta+2r^2} \tag{13b}$$

To compare the effects of roundoff noise in the block-floating-point realization to the effects in floating-point and fixed-point, we consider the input to be uniformly distributed white noise in the range

$$-\frac{1}{\sum_{n=o}^{\infty}|h_n|} < x_n < \frac{1}{\sum_{n=o}^{\infty}|h_n|}$$

where h_n is the filter impulse response. This then guarantees that the output will fit within a register. With these considerations, the normalized output noise-to-signal ratios for the first and second order filters are respectively

first order: $$\frac{\overline{\eta^2}}{\sigma_\epsilon^2\sigma_y^2} = \frac{1}{\sigma_y^2}\left[1+\frac{2k_1}{1-a^2}\right] \tag{14}$$

second order: $$\frac{\overline{\eta^2}}{\sigma_\epsilon^2 \sigma_y^2} = \frac{3}{G}\left[\frac{1}{\sin\theta}\sum_{n=o}^{\infty} r^n \left|\sin(n+1)\theta\right|\right]^2 + \frac{Gk_2}{\sigma_y^2}(2+2r^4+4r^2\cos^2\theta) \quad (15)$$

To compare (14) to the corresponding expressions for floating-point and fixed-point we will consider the high gain case and approximate A_n as given by (11) by $A_n \cong \frac{1}{2|y_n|}$. Assuming that y_n has a symmetric probability density about zero, we then have that $k_1 = 4\sigma_y^2$. Representing a_1 as $1-\delta$ with δ small we then approximate (14) by

$$\frac{\overline{\eta^2}}{\sigma_\epsilon^2 \sigma_y^2} \cong \frac{10}{\delta} \text{ (block-floating point)} \quad (16)$$

The corresponding approximations for floating-point and fixed-point are, respectively $(\frac{1}{\delta})$ and $\left(\frac{3}{\delta^2}\right)$. We observe, then, that for this high-gain approximation block-floating-point is approximately one bit worse than floating-point and, for the same size mantissas, better than fixed-point. Furthermore as $\delta \to o$ the noise-to-signal ratio for both floating-point and block-floating-point increase at a slower rate than fixed-point.

For the second order case we will restrict attention to a high gain filter (r close to one) and furthermore choose θ small enough to assume that $A_n \cong \frac{1}{2|y_n|}$ so that $k_2 \cong 4\sigma_y^2$. Again, letting $r = 1-\delta$, we introduce the high gain approximation $G \cong \frac{1}{4\delta \sin^2\theta}$. We can approximately bracket the expression

$$\frac{1}{\sin\theta}\sum_{n=o}^{\infty} r^n \left|\sin\left[(n+1)\theta\right]\right|$$

by noting that an upper bound is

$$\frac{1}{\sin\theta}\sum_{n=o}^{\infty} r^n = \frac{1}{(1-r)\sin\theta} \quad .$$

A lower bound is obtained by noting that the sum of the absolute values of an impulse response is the maximum attainable output value from a filter if the maximum input value is unity. Since the maximum output of the second order system at resonance is $\frac{1}{(1-r)(1+r^2-2r\cos 2\theta)^{1/2}}$, this provides a lower bound on the sum of the absolute values of the impulse response. For the high gain case this is approximately $\frac{1}{2\delta \sin\theta}$. Thus we will consider

$$\frac{1}{2\delta \sin\theta} \leq \frac{1}{\sin\theta} \sum_{n=o}^{\infty} r^n \left|\sin(n+1)\theta\right| \leq \frac{1}{\delta \sin\theta}$$

With these approximations, we have for the second order case that

$$\frac{1}{\delta}\left[3+\frac{8}{\sin^2\theta}\right] \leq \left(\frac{\overline{\eta^2}}{\sigma_\epsilon^2 \sigma_y^2}\right) \leq \frac{1}{\delta}\left[12+\frac{8}{\sin^2\theta}\right] \tag{17}$$

For comparison, the corresponding expressions for the floating-point and fixed-point cases are:

$$\left(\frac{\overline{\eta^2}}{\sigma_\epsilon^2 \sigma_y^2}\right)_{\text{floating-point}} = 1 + \frac{7}{4\delta \sin^2\theta} \tag{18}$$

and

$$\frac{6}{4\delta^2 \sin^2\theta} \leq \left(\frac{\overline{\eta^2}}{\sigma_\epsilon^2 \sigma_y^2}\right)_{\text{fixed-point}} \leq \frac{6}{\delta^2 \sin^2\theta} \tag{19}$$

Consequently as in the first-order case block-floating is only slightly worse than floating-point and better than fixed-point. Again, as $\delta \to o$ the noise-to-signal ratio for both floating-point and block-floating-point increase at a slower rate than fixed-point. An additional consideration is that (17), (18) and (19) compare noise-to-signal

ratios for equal size mantissas. Floating-point arithmetic requires additional bits in each word to represent the characteristic while block-floating-point requires additional bits to represent the characteristic for the entire block. Thus it is reasonable to speculate that in some cases for the same total number of bits per word, block-floating-point is the least noisy realization. While it is clear that the implementation of block-floating-point is more difficult than fixed-point it is almost certainly simpler than floating-point. Thus block-floating-point appears to warrant serious consideration as a means for implementing digital filters with hardware or on a digital computer with limited word size.

An additional consideration, is that, in block-floating-point final quantization of the input can be carried out just before the summer. If this is done, the variance of the output noise due to input quantization is reduced by a factor $\left(\frac{1}{A_n}\right)^2$.

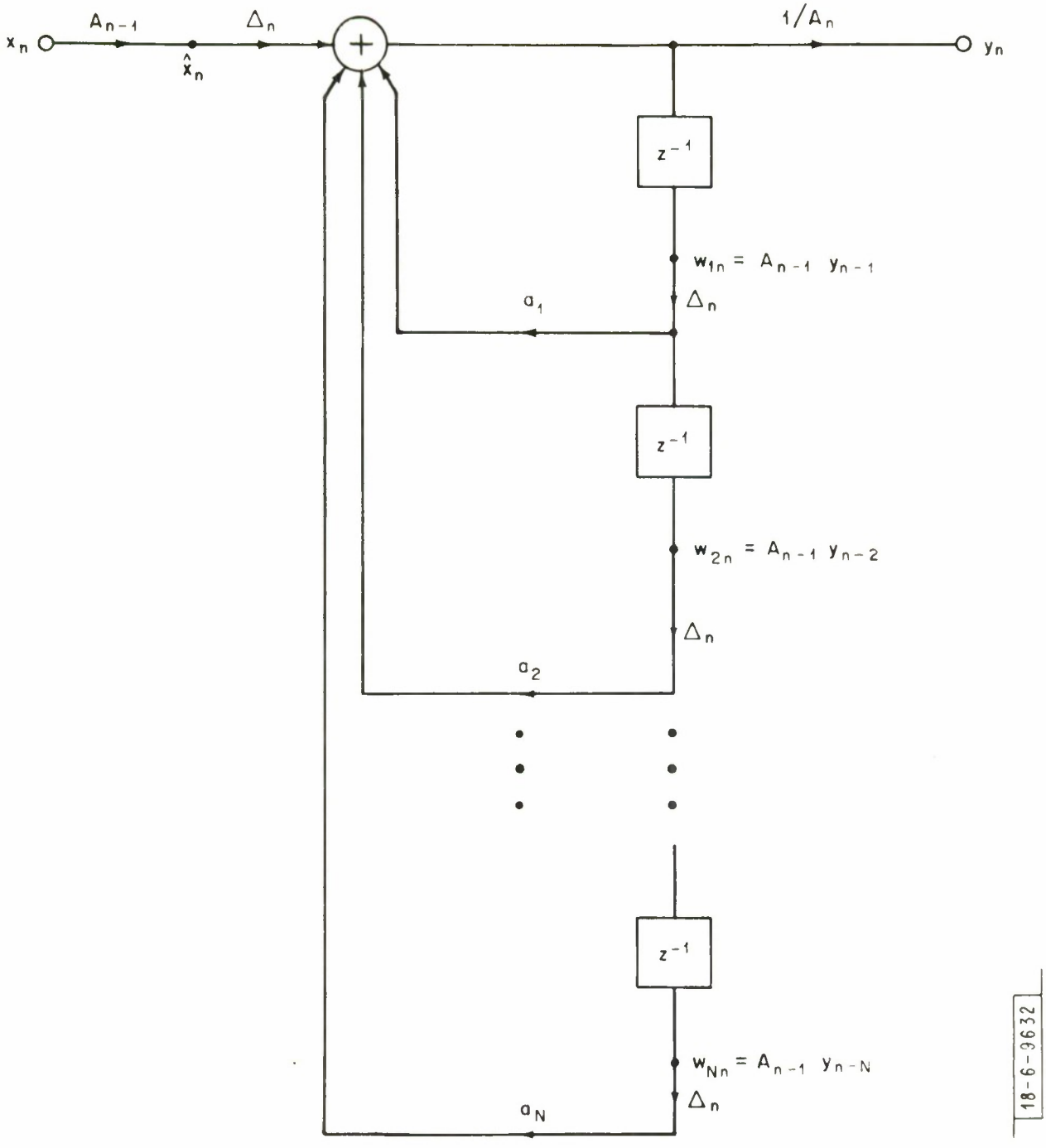

Fig. 1. Network for block-floating-point realization of an N^{th} order filter.

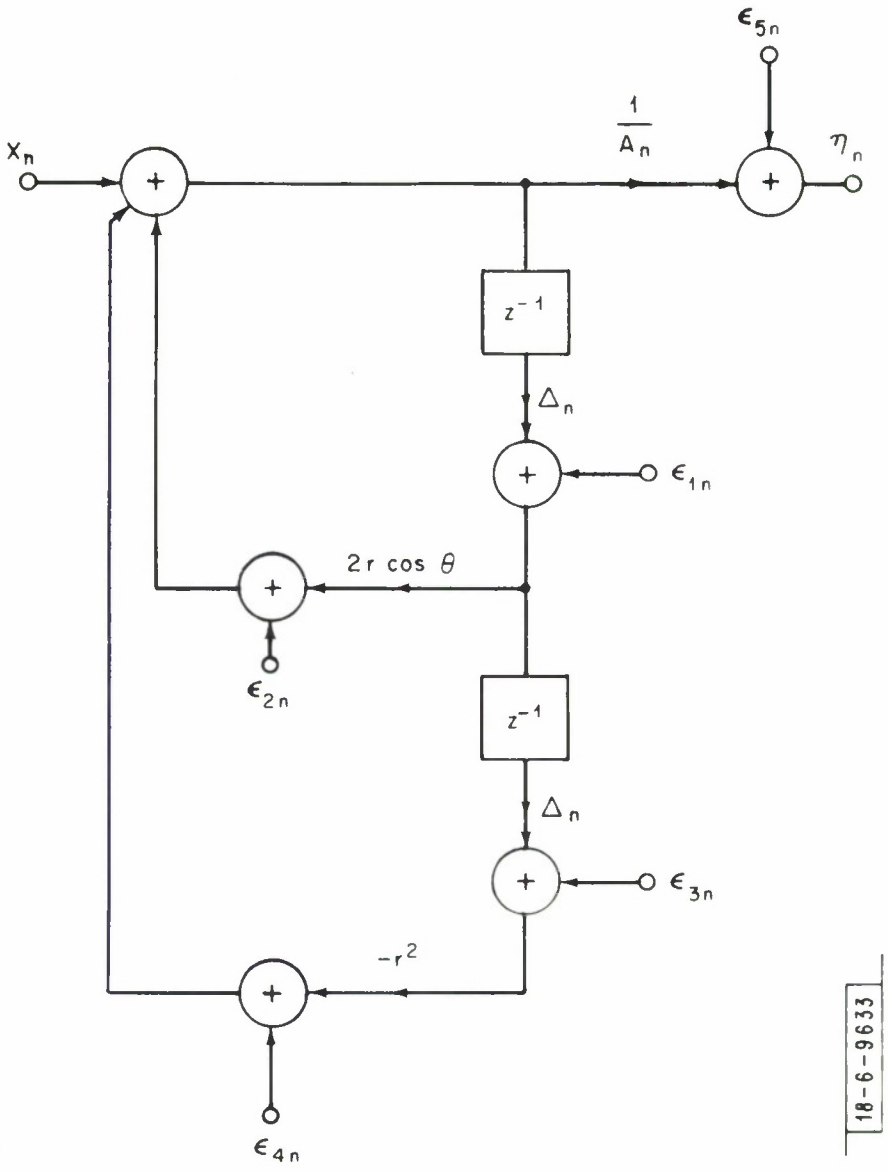

Fig. 2. Network for block-floating-point realization of a second order filter including roundoff noise sources.

REFERENCE

1. C. Weinstein and A.V. Oppenheim, "A Comparison of Roundoff Noise in Floating-Point and Fixed-Point Realizations of Digital Filters." (Submitted for publication)

DOCUMENT CONTROL DATA – R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY *(Corporate author)*
Lincoln Laboratory, M.I.T.

2a. REPORT SECURITY CLASSIFICATION
Unclassified

2b. GROUP
None

3. REPORT TITLE
Block-Floating-Point Realization of Digital Filters

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*
Technical Note

5. AUTHOR(S) *(Last name, first name, initial)*
Oppenheim, Alan V.

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
20 March 1969	16	1

8a. CONTRACT OR GRANT NO.
AF 19(628)-5167

b. PROJECT NO.
649L

c.

d.

9a. ORIGINATOR'S REPORT NUMBER(S)
Technical Note 1969-19

9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*
ESD-TR-69-63

10. AVAILABILITY/LIMITATION NOTICES
This document has been approved for public release and sale; its distribution is unlimited.

11. SUPPLEMENTARY NOTES
None

12. SPONSORING MILITARY ACTIVITY
Air Force Systems Command, USAF

13. ABSTRACT

A realization for digital filters using block-floating-point arithmetic is proposed. A statistical model for roundoff noise is presented and used to compare block-floating-point with fixed-point and floating-point realizations.

14. KEY WORDS
space communications
digital filters
fixed-point arithmetic
floating-point arithmetic